C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

The Evolution of a Portfolio Program: The Impact and Quality of the Vermont Program in Its Second Year (1992-93)

CSE Technical Report 385

Daniel Koretz, Brian Stecher,
Stephen Klein, and Daniel McCaffrey
RAND Institute on Education and Training

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

August 1994

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

CONTENTS

TABLES

SUMMARY

Since 1988, the Vermont Department of Education has been developing an innovative statewide performance assessment program. Although the program has several elements, it is best known nationally for its use of student portfolios in mathematics and writing in Grades 4 and 8. The program, which has been implemented statewide since the 1991-92 school year, was the nation's first effort to make portfolio assessment a cornerstone of an ongoing statewide assessment and has accordingly received widespread attention across the nation.

In 1990, RAND, as a partner institution in the Center for Research on Evaluation, Standards, and Student Testing (CRESST), has been evaluating the Vermont assessment program. RAND's evaluation, which is designed to provide feedback that will facilitate the program's evolution, has focused on three broad issues: the actual implementation of the program in schools and classrooms, the program's diverse effects, and the quality of the information yielded by the assessment. The evaluation has been focused specifically on the portfolio component of the assessment system.

This report presents results from the evaluation of the program in the 1992-93 school year. (For a comprehensive overview of the results from 1991-92, see Koretz, Stecher, Klein, McCaffrey, and Deibert, 1993.) It presents the results of interviews of principals in a stratified random sample of nearly 80 Vermont schools, questionnaires administered to mathematics teachers statewide, and analyses of the reliability of portfolio scores.

The results of the teacher questionnaire were in broad outline similar to those we obtained in 1991-92, which was the first year of full statewide implementation of the program. The 1991-92 questionnaires indicated that teachers perceive the program as causing substantial changes in mathematics instruction that are consistent with the goals of the program. Teachers indicated, however, that the program imposed substantial burdens on them, and they reported variations in program implementation that are substantial enough to threaten comparative interpretations of portfolio scores. These variations encompassed both the selection of tasks (e.g., their novelty and complexity) and the conditions under which they were performed (e.g., the amount of revisions students were permitted to make and the amount of help they were allowed to receive from parents and others). For the most part, the 1992-93 questionnaires do not reveal substantial changes in these patterns. In the view of teachers, the positive effects of the program on instruction continue, but the variations in implementation also remain unchanged.

Teachers did not clearly indicate a change in the time demands of the program; they did indicate that some activities, such as finding appropriate tasks for portfolios, were becoming easier. In 1992-93, teachers reported somewhat smaller discrepancies in student performance between portfolios and traditional mathematics activities.

In the spring of 1993, we interviewed principals in all but one of those schools in which we had interviewed principals the previous year, this time focusing our questions on changes in the portfolio program from the first year to the second. The principal interviews suggested somewhat more change than did the teacher questionnaires. In 1992, nearly half of the principals reported that the use of portfolios had been extended beyond the two subjects and grades required by the state program, which we interpreted as a strong signal of their positive evaluation of the program's value. A year later, more than 70% reported that portfolio use had been expanded beyond those two subjects and grades, and nearly all of the remaining principals anticipate such expansion in the future. However, a number of principals suggested that outside of the grades and subjects included in the state program, portfolios were not used for formal assessment.

Although some principals suggested that the portfolio program had become less burdensome to teachers between 1992 and 1993, the support they provided to teachers on behalf of the program, primarily in the form of release time, had not decreased. Indeed, nearly half of the principals commenting on change reported an increase in support, primarily because of a larger number of training sessions, workshops, and in-school meetings.

In the spring of 1992, the reliability of scoring of portfolios was low in both subjects and grades—sufficiently so to preclude most intended uses of the scores. In 1993, there was appreciable improvement in the reliability of scoring of mathematics portfolios. Expressed as correlations (which range from 0.00 when there is no relationship between the scores assigned by different raters to 1.00 when raters are in perfect agreement), the reliability of scoring at the level of individual scoring dimensions increased from roughly .40 in 1992 to .60 in 1993—still too low for many uses, but a clear improvement. When a single total score (summing over all 7 scoring dimensions) is created for each portfolio, the reliability of scoring increased from .60 and .53 in Grades 4 and 8, respectively, to .72 and .79. Although there is no simple standard for "how reliable is reliable enough," the scoring is reaching the point that reliability of scoring will no longer be the binding constraint for some uses of aggregate total scores. (It remains an impediment, however, for reporting at the level of individual dimensions and certainly for individual students.)

In contrast, the reliability of scoring in writing showed only trivial change from 1992 to 1993. For example, the reliability of scoring for total scores increased only from .49 and .60 in Grades 4 and 8, respectively, to .56 and .63.

In sum, the results from 1993 showed appreciable but inconsistent progress. Familiarity seems to be decreasing the burdensomeness of the program slowly, but time burdens and inconsistent implementation remain substantial concerns. In mathematics, considerable improvements in the reliability of scoring indicate the need to direct attention to other aspects of the quality of scores, such as other aspects of reliability and evidence pertaining to validity. The most discouraging finding is the low reliability with which writing portfolios were scored and the inconsequential progress made in this regard between 1992 and 1993.

CHAPTER 1: INTRODUCTION

Since 1988, Vermont has been developing an assessment program that is at the cutting edge of innovation in large-scale assessments. Although a rapidly growing number of statewide assessment programs incorporate some form of performance assessment, the Vermont program is unusual among them in that a centerpiece of the program is student portfolios and "best pieces" drawn from them. A pilot implementation of the program was conducted in 138 schools in the 1990-91 school year. The first statewide implementation of the assessment, in mathematics and writing in Grades 4 and 8, was conducted in the 1991-92 school year. The program has continued in the same subjects and grades since that time.

RAND has consulted with Vermont about the development and eventual evaluation of the assessment program since August 1988. Since 1990, RAND, as part of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), has been carrying out a multifaceted evaluation of the assessment program and its effects.

This monograph reports findings of the RAND/CRESST study in the 1992-93 school year. Additional findings about the reliability of the 1992-93 assessment and the quality of aggregate scores were reported earlier (Koretz, Klein, McCaffrey, & Stecher, 1993). (Detailed discussion of the findings from 1991-92 is presented in Koretz, Stecher, Klein, McCaffrey, and Deibert, 1993. For briefer presentation of results from both years and a discussion of their implications for policy, see Koretz, Stecher, Klein, and McCaffrey, in press.)

Background on the Vermont Assessment Program

Until recently, Vermont had no regular statewide assessment program. By the late 1980s, however, pressure was building to provide regular information on student performance, and by 1988, the state Department of Education began movement toward establishment of a statewide assessment system.

The deliberations that led to the decision to build the present, portfolio-based system are difficult to summarize succinctly because they were lengthy

and involved many diverse people, including the Commissioner of Education (Rick Mills), the Department's then-Director of Policy and Planning (Ross Brewer), the governor, members of the state board, local board members, teachers, and others. Several persistent themes, however, were stressed by Mills, Brewer, and others working to build the system. Ideally, the new system would:

- **avoid the distortions of educational practice that conventional test-based accountability appeared to have created in some other states;**
- **encourage good practice and be integrally related to the professional development of educators;**
- **reflect the Vermont tradition of local autonomy, "encourage local inventiveness, [and] preserve local variations in curriculum and approach to teaching" (Mills & Brewer, 1988, pp. 3, 5);**
- **provide "a high common standard of achievement for all students" (Mills & Brewer, 1988, p. 3); and**
- **encourage greater equity in educational opportunity.**

Those responsible for the nascent program were aware of the difficulties inherent in having an assessment program serve many functions at once and had been warned that some of their goals for the program pointed to different assessment designs. For example, a system designed to provide rich information about students and positive incentives for teachers might look very different from a system that was designed primarily to provide highly comparable information across schools.[1] The system that eventually emerged was intended to be a compromise among its many goals; for example, it should provide reasonable comparability across schools, but not at the cost of stifling good practice and local innovation.

The basic outline of the assessment program emerged quite quickly. Eventually, the assessment would span a broad range of subjects, but the state decided to begin with assessments in writing and mathematics in Grades 4 and 8. The assessment would have three components: year-long student portfolios, "best pieces" drawn from the portfolios, and state-sponsored "uniform tests."

---

[1] Daniel Koretz, presentation to the Commissioner Mills, Governor Kunin, and others, August 1988.

The details of the program, however, have been worked out only gradually. In contrast to the many states that either buy off-the-shelf tests or contract to have new tests built on a short schedule, the Vermont program was seen from the outset as a long-term and decentralized development effort. For example, in 1988, Mills called for mixing state-of-the-art assessment techniques with "emerging" techniques and warned that the development of the new program would be "a very long effort" (Mills & Brewer, 1988). Thus, in both subjects, the so-called "pilot" implementation in 1990-91 was less a true pilot of a developed program than an integral part of the development effort. Indeed, in mathematics, even the first full statewide implementation in the 1991-92 school year would be most accurately categorized as a combination of a developmental effort and a pilot test, rather than as an initial implementation of a fully planned program. Some of the details of the scoring of best pieces in the 1991-92 statewide implementation, for example, were not resolved until spring of 1992, and ratings of entire portfolios have not yet been attempted on a large scale.

Primary responsibility for the development of the portfolio and best-pieces components of the program was given to state-sponsored committees of teachers. These committees worked independently of each other, so the program evolved differently in writing and mathematics.

Mathematics

As implemented in 1992-93, the mathematics program required that students and teachers cull from each student's portfolio a set of five to seven "best pieces." The best-pieces sets of a sample of students from each participating classroom were sent to a central location for scoring by groups of volunteer teachers. All of the best pieces were graded on 4-point scales against seven criteria, four pertaining to problem solving and three to communication. The ratings on the individual pieces were then aggregated to provide an overall rating of the entire set of best pieces on each of the seven criteria.

The mathematics portfolio assessment was accompanied by the state's Uniform Test of mathematics. The UT is a matrix-sampled, mixed-format test, combining multiple-choice and open-ended items. Unlike the portfolio assessment, the UT was designed and scored by Insite, Vermont's testing contractor.

Writing

The design of the writing assessment is substantially different from that of the mathematics assessment. In writing, students' portfolios must include a set number of pieces of specified types, one of which is selected as the best piece. In Grade 4, each student's portfolio must include:

1. **a table of contents;**
2. **a single best piece, which is selected by the student, can come from any class and need not address an academic subject;**
3. **a letter explaining the composition and selection of the best piece;**
4. **a poem, short story, or personal narration;**
5. **a personal response to a book, event, current issue, mathematical problem, or scientific phenomenon;**
6. **a prose piece from any subject area other than English or language arts.**

The requirements for eighth grade are the same except that the portfolio must include three prose pieces.

The best piece and the rest of the portfolio were both scored on the same five dimensions:

- **Purpose**
- **Organization**
- **Details**
- **Voice/Tone**
- **Usage/mechanics/grammar**

A single 4-point scale is used with all five criteria. As in the case of mathematics, samples of portfolios were sent to a central location for scoring by volunteer teachers.

The writing portfolios were also accompanied by a Uniform Test of writing. This test was a direct writing assessment using a single prompt that was scored using the same criteria as were used with the portfolios.

Background on the RAND/CRESST Studies

The characteristics of the Vermont assessment program require that the RAND/CRESST evaluation be broad in scope. The RAND/CRESST evaluation is a series of interrelated efforts designed to gather information about:

- **the implementation and operation of the program at the school and classroom level;**
- **the quality of measurement (including reliability and validity); and**
- **effects on instruction and on other aspects of schooling.**

These questions have been addressed with a variety of methods, including questionnaires administered to teachers, interviews of teachers and principals, classroom observation, qualitative analysis of student portfolios, analysis of scoring methods and rubrics, questionnaires administered to scorers, and analysis of student-level and school-level scores.

The RAND/CRESST evaluation is formative. Our expectation, like that of the state Department of Education, is that the program will require a long period of development. Our evaluation is designed to monitor that process and to provide frequent corrective feedback along the way.

The Contents of This Report

Because of the state's need to use the results of the RAND/CRESST study for political decision making and program design, the results of the study are released piecemeal. Simple analyses of the reliability of portfolio scores and the quality of aggregate scores (for Supervisory Unions, which are groups of districts) were released in the fall of 1993 to facilitate the state's decisions about reporting (Koretz, Klein, McCaffrey, & Stecher, 1993). This report presents more elaborate analyses of the reliability of mathematics portfolio scores and adds the results of interviews with school principals and questionnaires administered to teachers.

CHAPTER 2: TEACHERS' PERSPECTIVES ON IMPLEMENTATION AND IMPACT

In the 1992-93 school year, as in previous years, we distributed questionnaires to mathematics teachers to explore their perceptions of the implementation and impact of the portfolio assessment program.[2] This chapter is based on data from questionnaires that were distributed to all teachers of mathematics in Grades 4 and 8 in the spring of 1993 along with the state's Uniform Test. The 1992-93 questionnaire addressed many of the same topics as previous surveys, but many of the items were revised or elaborated to provide more detailed information. (These changes preclude some direct comparisons with previous results.) Teachers were asked to complete the questionnaires anonymously and to return them with the completed student test booklets. Most survey questions were Likert-type items, requiring respondents to select one of five or six ordered responses. A few items required teachers to estimate the percentage of time devoted to particular activities or the percentage of students behaving in certain ways. There were two open-ended items requiring written responses.

Five hundred nineteen completed questionnaires were returned, three-fourths from Grade 4 and one-fourth from Grade 8. This represents approximately 52% of all Vermont teachers who taught mathematics in Grade 4 and 41% of mathematics teachers in Grade 8.[3] Although this response rate is much lower than last year's (83%), the total number of respondents is more than three times as large, including one-half of the entire population. Moreover, the characteristics of respondents, described below, suggest that they were reasonably representative of the total population of mathematics teachers in the target grades. A random sample of 50% of the

---

[2] This is the third year in which teacher questionnaires have been administered as part of the RAND evaluation; however, previous questionnaires were sent only to a sample of teachers. Much of this description of the results reported in this chapter was presented at the 1994 annual meeting of the American Educational Research Association (Stecher & Hamilton, 1994).

[3] These are the most conservative estimates of the response rate. They are based on the total number of teachers who teach mathematics at each grade level in the state. However, because of variations in distribution and testing procedures at the local level, we do not know that all eligible teachers received the survey.

papers with written responses to open-ended items was selected, and these responses were read, summarized, and tabulated by hand.

Results

This section begins with a description of the characteristics of teachers who completed the survey, then proceeds with a thematically organized discussion of the research findings. This discussion focuses on questions related to: *implementation* (specifically, in-service training and portfolio practices at the classroom level); *impact* (changes in curriculum and instruction, student performance, and teacher attitudes); and the *burdens* portfolios place on teachers and students.

Teacher Characteristics

The characteristics of teachers who responded to the survey are almost identical to those of the 1991-92 random sample, giving us more confidence in the generalizability of the survey results. The typical Vermont mathematics teacher has considerable classroom experience (see Table 2.1). On average, eighth-grade teachers have 16 years of experience and fourth-grade teachers a little less than 15 years. Less than 10% of the respondents have under 4 years of experience. Consistent with traditional elementary and middle school scheduling practices, 70% of eighth-grade teachers specialize in teaching mathematics (as opposed to teaching many subjects) while less than 2% of fourth-grade teachers specialize in mathematics.

Table 2.1

Characteristics of Sampled Teachers, 1991-92 and 1992-93

	1991-92		1992-93	
	Grade 4	Grade 8	Grade 4	Grade 8
Number	112	32	382	137
Response rate	90%	67%	52%	41%
Mean years' experience	15.0	16.7	14.6	16.2
Percent specializing in math	4.7%	73.0%	1.6%	69.3%

A large majority (82%) of respondents have at least one year's previous experience with the mathematics portfolios. For the most part, those teachers (18%) who had not used math portfolios before the 1992-93 school year simply did not teach mathematics at Grade 4 or Grade 8 in 1991-92. Over 20% of the teachers had two years experience with the math portfolio: One-fifth of the fourth-grade teachers and one-third of the eighth-grade teachers participated in the 1990-91 pilot program.

Training and Support

Vermont has provided portfolio-related training activities each year to meet the needs they perceived to be the greatest. For example, during the first year of implementation the focus of training was on explaining portfolio procedures and finding appropriate tasks. In 1992-93, the training focused on the scoring criteria. The overall level of satisfaction with training in 1992-93 was comparable to 1991-92. Over one-half of the teachers at both grade levels feel adequately prepared to work with the mathematics portfolios as a result of the training they received. Shortcomings in training were reported more often at Grade 4 than Grade 8. One-quarter of the fourth-grade teachers, compared to only 12% of the eighth-grade teachers, feel poorly or very poorly prepared to work with the portfolios. Fourth-grade teachers also rate the network scoring training sessions somewhat lower than do eighth-grade teachers. Approximately one-half of eighth-grade teachers rate the two network scoring training sessions as good or very good, compared to about 40% of the fourth-grade teachers.

Less than 10% of teachers wrote open-ended comments specifically about training. Most of these teachers say that training sessions placed too much emphasis on scoring portfolios and not enough attention was given to how to teach portfolios effectively. A few teachers complain of having to be away from their students too often to attend training sessions.

Variations in Classroom Implementation of Portfolios

One concern raised strongly in last year's RAND evaluation and echoed in teachers' open-ended comments this year was that portfolios are not implemented uniformly across classrooms and schools. Several items on the questionnaire reveal extensive variation in portfolio-related policies and

practices. For the most part, this variation has not lessened since 1991-92. For example, teachers' policies on revising best pieces still vary significantly.[4] Although 57% of teachers encourage revision of most best pieces, and 19% permit revision, another 19% *require* at least some revision, and 5% generally *do not permit* revisions. Similarly, the amount of time students spend revising varies widely. The average revising time is 30-40 minutes, but in roughly 17% of classrooms students do not revise at all. In another 15% of classrooms students take more than one full class period to revise a best piece. Students who are not encouraged or allowed to revise their best pieces will clearly be at a disadvantage relative to those who are encouraged, or even required, to revise their work.

There also is considerable variation in teachers' policies regarding who may assist students in revising their best pieces. One in four teachers does not assist his or her own students in revisions, and a similar proportion does not permit students to help each other. Seventy percent of fourth-grade teachers and 39% of eighth-grade teachers forbid parental or other outside assistance (see Table 2.2). The remaining teachers permit their students to receive outside assistance. This is consistent with 1991-92 results which indicated that 65% of teachers at Grade 4 and 43% at Grade 8 placed some limit on parental

Table 2.2

Assistance Allowed by Teachers on Best Pieces (Percentage of Teachers)

Source	Grade	Allowed to assist on which best pieces?				Rules differ for each student
		None	Some	Most	All	
The teacher	4	27	23	14	16	21
	8	27	32	9	13	19
Other students	4	34	31	11	12	11
	8	23	39	11	12	15
Parents or others outside of school	4[a]	71	13	4	4	8
	8	39	28	8	13	11

[a] Grade level difference significant at the 5% level ($p<.05$).

[4] The percentage of teachers reporting that students generally revise their best pieces at least once has risen from 73% to 80%. The mean number of revisions at Grade 4 is virtually unchanged from last year (1.17), but there is a modest increase from 1.00 to 1.10 at Grade 8.

assistance with portfolio projects. Further complicating matters, roughly 10% of teachers have different rules for each student. Teachers' policies also differ with respect to acknowledgment of outside help. Only about 20% require students to acknowledge or describe the assistance they receive, so a rater will not know whose work is represented on the page.

The type and quality of the work that becomes part of a student's portfolio is also heavily influenced by teachers' decisions about how best pieces are selected. Fourth-grade teachers generally provide students with more guidance in selecting best pieces than do eighth-grade teachers. However, this year's survey, like last year's, reveals substantial differences in the amount of teacher influence within grade levels, with some teachers playing an equal role with the student and others playing no role at all (see Table 2.3).

On the other hand, there are many similarities in portfolio practices. Most teachers are using mathematics portfolios with nearly all of their students: 96% reported that most, almost all or all of their students are compiling mathematics portfolios. Those who are excluded are primarily students from other grade levels who are enrolled in multigrade classes. About 15% of teachers also excuse some special education students from participation in the portfolio assessment.

Another area of congruity is in teachers' decisions about how much emphasis to place on different characteristics of best pieces. These decisions can have a subtle, but systematic, influence on the types of work students include in their portfolios. The vast majority of teachers place a moderate or heavy emphasis on the assessment scoring criteria and also on work that is

Table 2.3

Who Selects Best Pieces (Percentage of Teachers)

Who selects best pieces?	Grade 4[a]	Grade 8
Students on their own	21	30
Students with limited teacher input	55	57
Students and teachers have equal role	18	8
Teacher with limited student input	5	3
Teacher	1	1

[a] Grade level difference significant at the 5% level ($p<.05$).

"interesting or important to students." Most teachers place minor or moderate emphasis on students' pieces being mathematically correct and having a neat and polished appearance (see Table 2.4). Other than a small decrease in emphasis on student work being similar to examples in the Resource Guide, there has been very little change in emphasis since last year. It may be that training and scoring experience have helped to bring about this consistency of approach.

Changes in Curriculum and Instruction

One of the major goals of the mathematics portfolio program is to improve curriculum and instruction at the classroom level. In an attempt to measure these changes, we asked teachers to compare their current teaching activities with their approach before they started using portfolios.[5] As in 1991-92, most

Table 2.4

Teacher Emphasis on Portfolio Characteristics (Percentage of Teachers)

Area of emphasis	Grades	Amount of emphasis			
		None	Minor	Moderate	Heavy
Mathematically correct	4	5	32	54	10
	8	8	21	58	13
Neat and polished appearance	4	6	40	49	6
	8	9	38	49	4
Interesting or important to students	4[a]	1	6	52	41
	8	2	14	58	26
Similar to examples in Resource Guide	4	13	29	45	13
	8	17	32	42	9
Similar to good examples from scoring training	4	6	22	48	24
	8	10	25	46	19
Related to problem-solving criteria	4	1	6	42	51
	8	2	9	48	40
Related to mathematical communication criteria	4	2	10	47	41
	8	2	17	47	34

[a] Grade level difference significant at the 5% level ($p<.05$).

[5] The vast majority of teachers started using portfolios in 1991-92. However, approximately 20% of the teachers participated in the portfolio pilot the previous year. These teacher were comparing the present year to the year prior to 1990-91.

teachers report substantial changes in curriculum focus and teaching methods since they began using portfolios, changes that are consistent with the goals of the assessment program. These changes are more pronounced in the fourth grade than the eighth grade, which may be attributed to greater flexibility in scheduling and curriculum.

Curriculum changes are greatest in the areas of problem solving and mathematical communication, which are emphasized by the Vermont mathematics portfolio assessment. Most teachers are spending more classroom time in these areas in 1992-93 than they did prior to using portfolios. In the fourth grade, 83% of teachers devote more class time to "learning problem-solving techniques" than they did before the introduction of mathematics portfolios (see Table 2.5). Over 70% of fourth-grade teachers say they spend more class time applying math to novel and real world problems and solving logic or reasoning problems. Of the eight specific problem-solving

Table 2.5

Change in Time Spent on Problem-Solving Activities (Percentage of Teachers)

Activity	Grade	Somewhat or Much less	About the same	Somewhat or Much more
Exploring patterns	4	4	42	54
	8	5	57	38
Applying math knowledge to traditional word problems	4[a]	22	34	44
	8	17	56	28
Applying math knowledge to novel problems	4[a]	2	23	75
	8	1	29	70
Solving logic or reasoning problems	4[a]	1	24	75
	8	5	43	51
Applying math to problems in a real world setting	4[a]	2	26	71
	8	3	43	54
Collecting and analyzing data	4[a]	3	38	59
	8	10	45	44
Learning problem-solving techniques	4[a]	1	16	83
	8	3	34	63
Examining incorrect solutions	4	5	45	50
	8	8	54	38

[a] Grade level difference significant at the 5% level ($p<.05$).

activities mentioned in the survey, only traditional word problems are receiving the *same or less* class time in more than one-half of fourth-grade classes.

Curricular and instructional changes are not as great in the eighth grade. Fewer eighth-grade teachers report increases in class time devoted to problem solving than fourth-grade teachers. Less than one-half of the eighth-grade teachers spend more class time on four of the eight listed problem-solving activities. Nevertheless, two-thirds do give more attention to learning problem-solving techniques.

The changes are similar for mathematical communication. Eighty-nine percent of fourth-grade teachers and 77% of eighth-grade teachers are placing more emphasis on writing about math (see Table 2.6). Over 70% of fourth-grade teachers say they are devoting more time to explaining solutions to problems and discussing mathematics. But a substantially smaller percentage of eighth-grade teachers report such increases. A majority of eighth-grade teachers spend the same or less time in four of the five areas of mathematical communication listed on the survey.

Table 2.6

Change in Time Spent on Mathematical Communication (Percentage of Teachers)

Activity	Grade	Somewhat or Much less	About the same	Somewhat or Much more
Writing about mathematics	4[a]	3	8	89
	8	7	16	77
Explaining solutions to problems	4[a]	3	25	72
	8	15	53	32
Discussing mathematics	4[a]	1	29	71
	8	5	57	37
Making or interpreting charts, graphs, diagrams	4[a]	1	29	70
	8	7	45	49
Writing reports about mathematics	4	5	44	51
	8	9	41	50
Describing feelings about mathematics	4[a]	6	51	43
	8	11	60	30

[a] Grade level difference significant at the 5% level ($p<.05$).

There also have been changes in the types of instructional activities, although this has occurred less widely than changes in curricular focus. Just over two-thirds of teachers said that the portfolio assessment has moderately or greatly encouraged them to be innovative in planning mathematics lessons and activities. A slight majority of fourth-grade teachers engage in more open-ended activities and activities involving novel materials or supplies; but less than one-half of eighth-grade teachers do so (see Table 2.7).

The portfolio assessment also has affected the organization of mathematics instruction. About one-half of the fourth-grade teachers and one-third of the eighth-grade teachers have changed the way they group students during class in ways consistent with portfolio program objectives. There has been a modest shift away from individual work and toward whole class discussion, mixed ability groups, and working in pairs. However, for each type of class grouping, a substantial proportion of the teachers at both grades reported no change (see Table 2.8).

Increased attention to the topics and activities encouraged by the math portfolio program has come at a cost to other areas of the mathematics curriculum and, at Grade 4, to other subjects. Two-thirds of teachers are choosing to spend less time on computational skills and "other traditional

Table 2.7

Change in Classroom Activities (Percentage of Teachers)

Activity	Grade	Somewhat or Much less	About the same	Somewhat or Much more
Assign activities whose outcome and/or duration is uncertain	4	3	29	68
	8	2	36	63
Vary schedule or length of math activities	4[a]	1	27	72
	8	3	54	42
Involve students in hands-on math activities	4	3	46	51
	8	3	57	38
Use supplemental math books	4	14	50	36
	8	10	60	30
Use novel materials or supplies in math lessons	4	2	46	52
	8	3	50	47

[a] Grade level difference significant at the 5% level ($p<.05$).

Table 2.8

Change in Classroom Organization (Percentage of Teachers)

Activity	Grade	Somewhat or Much less	About the same	Somewhat or Much more
Discussing together as a whole class	4[a]	3	39	58
	8	14	58	28
Working in groups with students of similar ability	4	19	59	22
	8	18	59	23
Working in groups with students of different abilities	4	2	52	46
	8	2	62	36
Working in pairs	4	2	54	44
	8	3	60	37
Working individually	4	30	58	12
	8	32	64	3

[a] Grade level difference significant at the 5% level ($p<.05$).

math topics," and the majority of teachers agreed with the statement that "the portfolio assessment makes it more difficult to cover the mathematics curriculum" (see Table 2.9). In the fourth grade, 44% of teachers are spending less time on subjects other than math and writing.

Table 2.9

Changes in the Allocation of Class Time (Percentage of Teachers)

Activity	Grade	Somewhat or Much less	About the same	Somewhat or Much more
Any math activity	4[a]	15	39	46
	8	28	52	20
Computation	4[a]	65	30	5
	8	54	43	3
Other traditional math topics	4	63	32	5
	8	56	42	3
Any writing activity	4	6	31	63
	8	5	24	71
Subjects other than math and writing	4[a]	44	47	9
	8	9	71	20

[a] Grade level difference significant at the 5% level ($p<.05$).

We asked teachers to rate the frequency with which various classroom activities occurred. Problem-solving activities of one sort or another occur, on average, once per week (see Table 2.10). Although problem solving occurs less frequently than computation (which takes place two to three times per week), teachers indicate they are doing considerably more problem solving now than prior to the introduction of the portfolios. More unusual and challenging problem-solving activities occur less often.

Teachers expressed concern that "basic skills" are getting lost in the portfolio effort. In their written comments, they frequently noted that portfolio activities take time away from basic skills and computation, which still need attention. One of the most common open-ended comments was about the difficulty of finding time for the normal math curriculum and portfolios. As one teacher stated, "Until the curriculum outlines change to allow more

Table 2.10

Frequency of Class Engagement in Various Mathematics Activities (Percentage of Teachers)

Activity	Grade	Never	1-3 per Sem.	1-3 per Month	Once per Week	2-3 per Week	Daily
Computation and other traditional math topics	4	0	1	2	9	60	28
	8	0	2	6	12	51	29
Writing about mathematics	4[a]	2	9	21	41	24	3
	8	5	15	26	40	11	3
Applying math knowledge to solve novel problems	4	1	5	18	38	33	5
	8	2	9	35	33	17	4
Learning problem-solving techniques	4[a]	0	1	15	44	34	6
	8	1	5	27	28	29	11
Explaining solutions to problems	4[a]	0	2	11	41	34	11
	8	0	2	14	32	28	24
Working in groups with students of different abilities	4[a]	1	4	12	22	31	30
	8	5	8	21	18	24	24
Working on activities whose outcome and/or duration is unknown	4[a]	3	9	25	37	20	7
	8	4	17	32	31	13	3
Using novel materials or supplies in math lessons	4[a]	3	11	28	26	25	9
	8	3	22	27	29	12	7

[a] Grade level difference significant at the 5% level ($p<.05$).

portfolio-like tasks teachers will be doing a balancing act between covering the curriculum and embracing portfolio tasks." For many teachers, math portfolios are another add-on to an already busy curriculum, forcing them to make difficult choices.

Student Performance

Teachers are evenly split in their opinions about whether the program is promoting greater learning of mathematics. Fifty-one percent report that students are learning mathematics better because of the portfolios, while 40% believe student learning is "Neither better nor worse." Only 9% feel that portfolios have actually been detrimental to students.[6] We asked teachers to explain their responses to this item and 77% did so, often in considerable detail. Positive statements about student learning (made in 69% of the comments) focus mainly on improvements in students' thinking and reasoning about math. Also common are comments that portfolios encourage students to explain their ideas and relate math to real life, which improves their understanding of mathematical concepts.

Over one-half of the teachers made negative comments about the impact of portfolios on student learning, often mentioning that learning is worse (or not any better) because other areas of the math curriculum have to be cut to make time for portfolios. The most frequent negative teacher remark (made by 15% of the teachers who commented on this item) is a reference to cutting back on basic skills or computation. Many feel the need for better balance between these activities and portfolios. Another frequent concern is that younger students are being turned off to math because of the writing demands of portfolio tasks. Several fourth-grade teachers (11% of the student learning comments) mentioned that the writing required for math portfolio tasks is developmentally inappropriate, particularly writing that relates to the PS4 criterion.[7] Teachers also repeatedly expressed the need for portfolios to be implemented at all grade levels for there to be a significant impact on student

---

[6] These results are based on teachers' professional judgment about student learning and student ability levels. They should be interpreted cautiously since student learning may be difficult to characterize across ability levels, especially in the midst of a substantial shift in curriculum focus and instructional practice.

[7] PS4: What decisions, findings, conclusions, observations, connections and generalizations has the student reached?

learning. Three-quarters of teachers at Grade 4 and two-thirds at Grade 8 agree with the statement that "Math portfolios should be expanded to all students in all grades."

Differences between students' performance on traditional mathematics assignments and portfolio tasks were less this year than in 1991-92. On average teachers said one-half of their students performed about the same on the two types of tasks, compared to about one-third of the students in the previous evaluation. Nevertheless, about one-third of fourth-grade students and one-quarter of eighth-grade students did worse on portfolio tasks than on traditional math assignments, while the remainder did better.

Teachers report a relationship between students' ability levels and how well they respond to portfolio work. Teachers generally think that high-ability students have a more positive reaction to portfolios than do low- and average-ability students. For example, high-ability students are more likely to "enjoy portfolio work more than regular math assignments," and are less likely to be hampered on math portfolio tasks because of poor writing skills (see Table 2.11). Teachers indicate that a smaller proportion of their low-ability students are "learning more math because of portfolios." And while most teachers report that few or none of their students find portfolio problems easier than traditional assignments, they find this to be true least often with low-ability students.

Teacher Attitudes Toward the Portfolios

Teachers have mixed views about the mathematics portfolio program. Although there is broad support for portfolios, there is also substantial concern about the implementation of the program and about specific uses of portfolios. Teachers' written comments reflect a mix of enthusiasm and frustration over portfolios. Statements of support for the philosophy behind portfolios are often followed by concerns about state demands.

Teachers think the portfolios are helpful as informal classroom assessment tools but worry about their use for external assessment purposes. The majority agree or strongly agree that the portfolios help students monitor their own progress, and that portfolios are useful for informing parents about student progress. The majority also agree that portfolio scores should be used

Table 2.11

Student Reactions to Mathematics Portfolios by Grade and Ability Level (Percentage of Teachers)

Student reactions	Grade	Percent of teachers reporting Most/Almost all		
		Low-ability students	Ave.-ability students	High-ability students
Enjoy doing portfolio tasks more than regular math assignments	4	10	15	43
	8	18	23	38
Like portfolios better this year than last year	4	14	21	36
	8	17	23	28
Learn more math because of the portfolios	4	21	30	49
	8	25	28	32
Find portfolio tasks easier than traditional assignments	4	4	6	14
	8	10	10	17
Portfolio tasks do not reflect math ability because of poor writing skills	4	46	5	3
	8	31	6	2

as part of students' grades, although about one-half of the teachers judge students' math work differently when assigning grades than when scoring for the portfolios. Most find the portfolio criteria easy to use, but about one-third report frequent difficulty applying criteria PS3, PS4 and C1 (see Table 2.12).

In contrast, teachers are more cautious about the use of portfolios for external assessment purposes. The vast majority of teachers do not believe it would be fair to evaluate students on the basis of their portfolio scores. While the majority of fourth-grade teachers think portfolio scores are a better measure of math learning than standardized tests, eighth-grade teachers are about evenly divided between those who agree, those who disagree, and those who are uncertain. One of the most common concerns raised by teachers in their open-ended comments was the state's strong emphasis on scoring. Many feel that the emphasis on reliable scoring is misguided and perverts the original purpose of portfolios as a tool for assessing an individual student's growth. One teacher noted that "the state wants portfolios to be scored like a bubble test for their own purposes . . . Theoretically, the portfolios were to show personal growth in math and writing abilities."

Table 2.12

Difficulty Applying Scoring Criteria to Student Portfolios (Percent of Teachers)

Criteria	Grade	Never or Seldom	Occasion-ally	Often or Very often
PS1 Understanding	4	76	20	4
	8	73	24	3
PS2 How?	4	52	38	10
	8	51	43	6
PS3 Why?	4	24	48	28
	8	27	47	26
PS4 What?	4	30	32	38
	8	35	25	40
C1 Language	4	32	40	28
	8	32	38	30
C2 Representations	4	44	41	15
	8	47	41	12
C3 Presentation	4	43	40	16
	8	41	47	12

Teachers are concerned about the validity of portfolios as an assessment instrument. Expressing a common sentiment, one teacher asked, "How can the validity of scoring outside the classroom be justified when there are so many uncontrolled variables?" About one out of every four (23%) teachers who commented on "other issues" expressed concerns along these lines. Several teachers (5%) also worried that some of their colleagues were providing students with opportunities to improve their work that they felt were inappropriate and were not permitted in their own classes.

There is a strong sense that the changes brought about by portfolios have had a positive impact on mathematics education. Many teachers expressed some support for the portfolio philosophy along with their complaints about the program. For example, 54% of teachers agreed with the statement that "the Vermont mathematics portfolio assessment is moving education in the right direction," while only 21% disagreed. Seventy-five percent of the teachers also supported expanding the portfolio assessment to all students in all grades.

Time Burdens

As in the past, time burdens are teachers' greatest concern; portfolios consume considerable time both in class and outside of class. However, the questionnaires offered inconsistent information on the extent to which the time burdens of the portfolio program have changed. In 1993, we asked teachers both to estimate the hours they devoted to specific portfolio-related activities and to compare the time burden to the previous year. On the one hand, 55% of teachers who used portfolios in 1991-92 report that they are spending more out-of-class time this year on portfolios than they did last year, and 60% are spending more classroom time than in 1991-92. Fewer than 10% report decreases compared to either year. On the other hand, teachers' reports of the actual number of hours spent in classroom and other portfolio-related activities have gone down by approximately one-third from 1991-92. The discrepancy between these two estimates of change may be partly explained by changes in the way the absolute time estimates were gathered between the two years or by unmeasured nonrepresentativeness of the 1993 respondents.[8] Since the vast majority of teachers reported spending the same or more time this year than last, we place more confidence in these comparative judgments than in the estimates of specific hours, which do show a decline in time burden.

Other responses support the conclusion that the portfolios continue to make significant demands on teachers. For example, most teachers at both Grade 4 and Grade 8 feel that they spend too much time managing and scoring portfolios (see Table 2.13). Most teachers do not feel that the demands of the mathematics portfolio program are lessening, and many are displeased that the burden continues to be so great. Less than one-third of the teachers agreed with the statement that "Overall, portfolios are less of a burden on me this year than last year."

Many observers expected that the demands placed on teachers by the portfolio assessment program would diminish as teachers became more experienced, and this was true to some degree. Roughly 40% to 50% of the teachers who participated in the program in 1991-92 said that specific portfolio-

---

[8] That is, differences in format between the 1992-93 and 1991-92 questionnaires may have contributed to this inconsistency.

Table 2.13

Demands of the Mathematics Portfolio Program on Teachers (Percentage of Teachers)

Statement	Grade	Disagree	Neutral	Agree
It is easy to prepare portfolio lessons	4	60	22	18
	8	52	30	18
I spend too much time managing portfolios	4	21	26	53
	8	22	28	49
Overall, portfolios are less of a burden on me this year than last year	4	55	16	29
	8	52	25	22
Scoring portfolio work is not too time consuming	4	84	7	9
	8	83	7	10
The portfolio assessment makes it more difficult to cover the mathematics curriculum	4	21	15	65
	8	16	20	64

related activities, such as finding interesting tasks and teaching problem solving, had become easier by 1992-93. However, in most cases, a similar proportion of teachers said that the activity was no easier than the year before (see Table 2.14). Similarly, we could find little evidence that teachers with three years of portfolio experience found portfolios easier or less time consuming than teachers in their second year. Third-year portfolio users were slightly less likely than second-year portfolio users to find managing portfolios too time consuming (54% to 47%); but this was the only significant difference.

One of the most common issues raised by teachers in their open-ended comments (mentioned by 25% of those who commented on "other issues") was the excessive burden placed on fourth- and eighth-grade students and teachers by the combination of writing and mathematics portfolios. They feel they spend too much time away from their students for training and that they carry a burden that is not placed on teachers and students at other grade levels. One teacher warned, "I support portfolios but fear that you'll lose your allies (even me) by over working them!"

Table 2.14

Change in Difficulty of Portfolio-Related Teacher Activities (Percentage of Teachers)

Activity	Grade	More difficult	About the same	Easier
Find interesting tasks	4	10	41	49
	8	17	47	36
Decide if task is appropriate for portfolio assessment	4	9	35	56
	8	15	32	53
Integrate tasks into the math curriculum	4	13	44	43
	8	21	48	32
Teaching problem solving	4	10	39	51
	8	6	48	46
Teaching mathematical communication	4	16	46	39
	8	14	47	40
Motivate students to work on portfolio tasks	4	18	52	30
	8	26	46	28
Make students understand qualities of good pieces	4	18	41	42
	8	19	34	47
Explain the portfolios to parents	4	9	56	36
	8	6	68	26

Summary

This year's results are similar to last year's. The 1991-92 teacher survey revealed significant changes in curriculum and instructional practices consistent with the goals of the portfolio assessment, but also indicated that portfolios placed a substantial burden on teachers' time. Moreover, there were significant variations in teachers' approaches to portfolios, which would affect the interpretation of portfolio scores.

This year's questionnaire reveals very little change between 1991-92 and 1992-93. The level of satisfaction with training in 1992-93 was comparable to 1991-92. For the most part, variation in portfolio-related policies and practices has not lessened since 1991-92. For example, substantial differences remain in teachers' revision policies, in the amount of teacher influence on the choice of best pieces, and on the limit on parental assistance with portfolio projects. On a more positive note, desired changes in curriculum and instructional focus

were sustained in the second year, and teachers reported lesser differences between students' performance on traditional mathematics assignments and portfolio tasks than in 1991-92. Some activities, such as finding tasks, were becoming easier, but there was no substantial reduction in the time burden portfolios placed on teachers.

Overall, the message from Vermont teachers about the portfolio assessment system remains mixed. Most teachers have modified their curricula and teaching practices to emphasize problem solving and mathematical communication skills, but many feel they are doing this at the expense of other areas of the curriculum, especially basic skills and computation. About one-half of the teachers see a payoff for their extra effort in terms of improved student learning; one-half do not. Most teachers express support for mathematics portfolios in a general sense, but there are widespread concerns about using portfolios as an external evaluation tool and, most of all, about the time demands of planning, administering, and scoring portfolio problems. Furthermore, it is clear that variations in teachers' approaches to implementing mathematics portfolios persist.

Teachers' responses suggest some ways Vermont might improve the portfolio assessment system in the future. First, teachers express strong support for expanding portfolios to all grade levels. For many teachers this is point of efficiency as well as fairness. As long as portfolios are limited to Grades 4 and 8, other grades will be slow to adopt practices that support the skills emphasized by the portfolio assessment system. Also, parents, administrators, and other teachers will continue to expect all teachers at Grades 4 and 8 to abide by the "traditional curriculum." Secondly, the developmental appropriateness of certain aspects of the Grade 4 mathematics portfolios should be re-examined. Many fourth-grade teachers are convinced that the writing demands are too great for many of their students. Finally, the state Department of Education should review the balance between local flexibility and standardization of implementation—while flexibility contributes to the meaningfulness of the portfolios as local instructional tools, it reduces the validity of inferences that can be drawn from the portfolio scores.

CHAPTER 3: PRINCIPALS' VIEWS OF THE PORTFOLIO PROGRAM

In the spring of 1992, during the first year of statewide implementation, we interviewed a representative sample of 77 fourth- and eighth-grade principals about their experiences with the portfolio assessment program. The schools were a stratified random sample (stratified on estimated poverty rates and school size) of population of Vermont schools containing fourth or eighth grades. In the spring of 1993, during the second year of the program, we returned to the same schools to re-interview the principals in an effort to gauge changes in their experiences over the intervening year. We obtained interviews from 76 principals, 38 in each grade. The results of the second round of interviews of principals are reported here.

Expanded Use of Portfolios

Nearly half of the principals we interviewed during the first full year of program implementation (1991-92) reported that the use of portfolios in their schools had been expanded in some manner beyond the two subjects and two grades included in the state assessment program. We considered this voluntary expansion of portfolio use a clear sign that educators considered them a worthwhile approach, despite the many substantial burdens they impose. When we re-interviewed the principals of the same schools a year later, we investigated whether the use of portfolios had continued to expand, and we explored the nature of the additional uses of portfolios.

The use of portfolios continued to expand during the second year of statewide implementation. Just under half of the principals reported additional use of portfolios in 1992; more than 70% reported additional use by the spring of 1993.[9] All but one of the principals who reported additional use of portfolios clarified that they were used in at least one grade other than the two (fourth and eighth) included in the state's program.[10] About three-fourths

---

[9] Two secondary school principals appeared not to know whether additional teachers were using portfolios and were deleted from calculation of this and the following percentages.

[10] Two principals of small schools indicated that the use of portfolios in other grades stemmed from the fact that their fourth-grade students were in mixed-grade classrooms. This does not appear to be a primary cause of the expansion statewide, however. It was not mentioned by

reported that the expansion was only in mathematics and writing, and 9% specifically reported expansion in other subjects as well as mathematics and writing.[11] We did not probe why principals expanded primarily in mathematics and writing. It is possible that it was simply easier because the program was already developed and functioning in those subject areas. However, it may be that part of the explanation rests with a widespread view that portfolios are valuable for tracing students' growth longitudinally. In response to a number of our questions, 43% of the principals commented that portfolios are valuable for tracking progress longitudinally, even though this function is not incorporated into the state's portfolio program.

Our interviews suggest that educators will continue to extend the use of portfolios. Almost all (90%) of our respondents who had not yet extended portfolio use said they expected to in the future. Indeed, two-thirds of them reported having already made concrete plans for doing so. In addition, about 80% of those who reported having already expanded portfolio use anticipate yet further extension in the future. Many of the principals expect that the use of portfolios will be extended in the future to other subject areas. Science, social studies, and art were often mentioned as areas for future portfolio use. A few principals (about 9%) specifically reported the goal of eventually using portfolios for interdisciplinary instruction.

An appreciable number of principals, however—about 17%—expressed some uncertainty about future expansion because they did not anticipate that the decision would be theirs. Seven of the 76 principals reported that future expansion would be primarily determined by decisions by district or state officials, and 6 said that it would be primarily determined by teachers.

Although we did not explicitly ask about the uses to which school staff put portfolios outside of mathematics and writing in Grades 4 and 8, a fourth of the principals volunteered comments. Their comments suggest that portfolios collected at the initiative of local educators are used in ways very different from

---

other elementary school principals. Moreover, split-grade classes stemming from small school size are less likely in secondary schools (which are larger in Vermont), and if those two cases are taken into account, the percentages of secondary and elementary principals reporting expansion were the same.

[11] One principal reported expansion but provided too little detail to clarify its nature. Eight reported expansion in mathematics or writing but did not specify whether they had also expanded in other subject areas.

the state's use of portfolios. Two-thirds of the 19 principals who volunteered comments of this sort said that in their schools, portfolios collected on their own initiative were not used for formal assessment. However, we do not know how representative this small number of principals is.

Changes in Burden

Of the principals interviewed in 1992, 86% labeled the portfolio program as burdensome, and some of the others noted specific burdens despite not characterizing the program that way. More often than not, principals pointed to teachers rather than themselves as the people most substantially burdened by the portfolio program. Consistent with these comments by principals, mathematics teachers responding to our questionnaire in 1992 reported that the program created large time demands for them and imposed other stresses as well, such as difficulties finding appropriate tasks and uncertainty about expected procedures (see Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). Some of the specific burdens noted by teachers, however, should have declined as teachers became more familiar with the program and accumulated the knowledge and materials required. Accordingly, when we re-interviewed principals in 1993, we asked for their opinions about changes in the burden imposed by the program.

Sixty-one principals provided responses that referred unambiguously to changes in the burdens imposed by the program between 1992 and 1993.[12] About half of these (31 of the 61) reported a decrease in burden. Twenty-one percent of these principals reported an increase in burden; about an equal number reported no change; and 8% pointed out both increases and decreases in burdens. Recall that of the mathematics teachers responding to our questionnaire, 40% to 50% said that specific portfolio-related tasks had become easier (Chapter 2).

Some clear patterns appeared in principals' explanations of changing burden. Of the principals who reported a decrease in burden, about half pointed to increased familiarity as one reason for the improvement. For example, one elementary school principal noted, "The program has become

---

[12] Unfortunately, 15 additional principals provided responses that enumerated burdens but did not clearly enough focus on change for us to include them in these tallies.

less burdensome for all of us. Having had put it into practice last year, having been through all the ropes, all is easier and more enjoyable to do." Of the 13 principals that reported an increase in burden, 9 referred to the time requirements of the program.

The decrease in burden experienced in some schools did not necessarily indicate that operation of the program was no longer burdensome. Indeed, 15 of the 31 principals who reported a decrease in burden noted aspects of the program that remain burdensome; 13 of these made specific comments about time demands, the difficulty of assimilating new teachers into the program, or frustration with the state's management of the program. Moreover, one type of cost—support for teachers, described in the following—appears to have held constant or increased.

Special Support for Teachers

All but one of the interviewed principals reported that they have provided special support to teachers for portfolio-related work. Release time was almost universally mentioned as a primary form of assistance to teachers. Three-fourths of principals specifically noted providing release time for workshops and training sessions; 30% reported release time for scoring; 21% reported release time for meetings; and about half reported release time for other portfolio-related duties. (Scoring for state reporting was carried out after school was out for the year, so principals were presumably referring to scoring for internal purposes, which the state encourages.) Over 30% of principals noted that they provided teachers with as much release time as they needed for portfolio-related training and workshops.

Seventy-one percent of the principals specifically addressed the issue of changes in the support they provided to teachers.[13] Of those, only a few (3 of 54 principals, or 6% of those who specifically addressed change) reported offering less support for teachers in 1992-93 than in the previous year. These respondents asserted that teachers were more comfortable with the system

---

[13] Our interview protocol specifically asked about changes, but at the end of a series of three questions about teacher support. Analysis of transcripts suggests that at least one interviewer either did not prompt sufficiently when principals discussed support but not change or failed to make sufficient note of their comments about change. Therefore, the discussion that follows focuses on the 54 of 76 principals whose transcripts unambiguously discuss either change in support or the lack thereof.

and therefore did not require as much release time for training and other purposes. The largest group (27 of 54 principals, or 50%) reported no change in the level of support they provided to teachers.

Twenty-four principals (44% of those discussing change) reported an *increase* in the amount of support offered to teachers. The increase was attributed primarily to a greater number of training sessions, workshops, and in-school meetings, and more frequent release time for scoring and other portfolio tasks. Only two principals attributed the increase in support to the broadening participation in the portfolio program by teachers in other subjects and grade levels. Interestingly, fourth-grade principals were two-and-a-half times as likely as eighth-grade principals to report increasing the level of special support offered to teachers over the past year.

Teacher Attitudes

We polled principals about their perceptions of teachers' attitudes toward the program. Only a minority reported changes in teachers' attitudes since 1992, but of the 34 who did, 25 (74%) reported more positive attitudes, and only 9 reported that attitudes remained largely unchanged. Not a single principal reported increasingly negative attitudes. Consistent with these reports of change, 54% of the total sample of the principals interviewed in 1993 characterized their teachers' attitudes as predominantly positive, in contrast to 23% interviewed the year before. As noted in the previous chapter, mathematics teachers' responses to our questionnaire were mixed, showing both enthusiasm and concerns.

Student Attitudes

Principals typically have far less contact with students than do teachers. However, the majority of Vermont schools are small enough that we believed principals would have some basis for an opinion of students' attitudes toward the portfolio program, and we asked them to characterize students' attitudes in our interviews in both 1992 and 1993.

In 1992, only a third of the principals characterized their students' attitudes toward the program as clearly positive; by 1993, more than half (55%) did. Twenty-three percent of the principals interviewed in 1993 reported that

students' attitudes were mixed, and only 5% reported predominantly negative attitudes. An additional 8% of principals reported not knowing their students' attitudes or provided other unusable responses. Interestingly, all of the principals reporting predominantly negative attitudes were reporting on eighth-grade students.

Parents' Views

In 1993, as in 1992, principals reported received limited feedback from parents about the portfolio program. More than half of the principals interviewed in 1993 reported receiving no feedback from parents, and an additional 19% discussed parental reactions but inferred them from indirect evidence. Only 26% of the principals reported direct evidence of feedback from parents. Moreover, some principals based their views on comments by very few parents.

Of those principals who offered an opinion of parental views, 33% characterized their views as positive, while about 15% described their views as negative. The remaining principals characterized parental views as neutral or mixed. Both positive and negative comments were diverse. About one-fourth of the principals offering an opinion of parental views (7 of 33) noted concerns about lessened emphasis on traditional instruction and basic skills. However, given the small numbers of principals who reported parental views and the apparently meager evidence on which many based their opinion, we would not place much confidence in their descriptions.

Changes in Instruction

In the spring of 1993, as the second year of statewide implementation was drawing to a close, nearly three-fourths of the interviewed principals reported that the portfolio program has produced positive changes in instruction. Another 17% reported a mix of both positive and negative effects. Only a few principals reported either primarily negative effects (2 principals, or 3%) or that they were not aware of effects (6 principals, or 8%). These responses are somewhat more positive than the overall assessments of mathematics teachers responding to our questionnaire, 54% of whom said that the portfolio program is moving instruction in the right direction. The principals' responses, however, are more consistent with the specific information

teachers provided about instructional change. For example, between 70% and 89% of teachers reported more discussion of mathematics, explanation of solutions, and writing about mathematics (Table 2.6).

These comments represent a moderate change from the 1992 school year. In 1992, a fourth of the principals felt it was to early to make a judgment about instructional effects. The proportion offering clearly positive appraisals was modestly lower than in 1993 (60% versus 72%), and the percentage offering mixed views was much lower (3% versus 17%).

The positive effects noted by principals were diverse. We attempted to place them in two categories: (a) greater emphasis on problem solving and other higher order thinking skills; and (b) other changes in curriculum (e.g., interdisciplinary lessons) and instructional style (e.g., greater use of projects and group work). This division is imprecise and somewhat arbitrary. For example, one could classify an increase in the use of writing in mathematics lessons as either a curricular change or an emphasis on higher order thinking. However, only a modest number of responses were ambiguous in this regard, and different decisions about classification had only minor effects on the percentages reported here. We found that about 45% of principals made specific reference to an increased emphasis on higher order thinking skills, most often specifically problem solving. Nearly 70% made reference to one or more other curricular or instructional changes. This latter group of comments was too diverse to classify clearly. It included, for example, lessened reliance on textbooks and worksheets; an increase in writing overall and more integration of writing with other subjects; more work in cooperative groups; a greater use of real world problems; and greater clarity and consistency in instructional goals. It is important to emphasize, however, that each of these comments was typically offered by only a few principals.

Principals' comments about negative effects, although infrequent, were relatively consistent, focusing on a perceived excessive emphasis on portfolio-related work to the detriment of other parts of the curriculum. One principal specifically noted that one effect had been a decrease in scores on "mechanical" aspects of mathematics on their norm-referenced measure.

Impact on Student Learning

Principals were asked whether they had observations about the impact of the portfolio program on student learning. Overall, their responses indicated that many principals knew little or nothing about these effects. About one-third (or 25 of 78) of the principals explicitly indicated that they did not know what the impact of the program had been or were so clearly just speculating that we treated their answers as indicating no information. Eighth-grade principals were two-and-a-half times as likely (47% versus 18%) to indicate outright that they did not know. Some principals offered clear statements about effects but seemed to be basing them on relatively little information. (A few carefully drew this distinction for us, reporting their observations but qualifying them by saying that firm data—e.g., test scores—were not yet available.) Moreover, when asked to explain the ways in which students' learning had changed, many principals referred to learning-related attitudes and behaviors rather than, or in addition to, actual learning.

Accordingly, the principals' responses to this question must be interpreted cautiously, but the patterns within them were consistent enough to warrant discussion nonetheless. Most principals (55%) said that the effects were positive. Twice as many fourth-grade principals as eighth-grade principals reported positive effects (73% versus 37%). A much smaller number of respondents reported mixed effects (5%) or no real change in student learning (6%). Only a single principal reported that the effects of the program on learning were primarily negative.

Changes in aspects of cognition, learning, and students' academic performance were the most frequently cited effects of the program. These were reported by 60% (32 of 53) of the principals who reported that they knew what the impact of the program had been (41% of all principals). Of the 32 principals who reported such changes, 14 (43%) cited changes in critical thinking, reasoning, depth of thinking, and the like. Seven principals (22%) noted transfer of portfolio-related skills to other subject areas, and 8 principals (25%) referred to some aspect of communication or articulation of ideas (in some instances, specifically the use of mathematical language).

Utility as Internal and External Assessment

In 1993, principals remained relatively positive about the use of portfolios for internal assessment but negative about their use externally, for example, for comparing schools. In this respect, their opinions parallel those of teachers (see Chapter 2). More than half (57%) of the principals stated that portfolios are useful as internal assessment, although a sizable number of these principals also cautioned against using portfolios in isolation. Interestingly, 20% of all principals reported that portfolios are useful as internal assessment because of their ability to chart students' progress over time, even though the state's implementation of portfolios itself provides no measure of student growth and no ready means of assessing it. However, even as the second year of statewide implementation was ending, 25% of the principals stated that it was still too early to use portfolios as assessment tools. Many of these specified that it was even too early to be confident of the utility of portfolios as internal assessments.

Well over half (59%) of principals opposed the use of portfolios for comparing schools. Thirty-seven percent offered mixed views about this use of portfolios, and only a single principal was clearly supportive of that use. (Two principals had no clear opinion.)

Principals' objections to the use of portfolios for external comparisons were diverse. Thirty-eight percent of all interviewed principals expressed concerns that the validity of comparisons among schools would be undermined by noneducational differences among them, such as differences in student characteristics or available resources. About one-third (34%) were concerned about the impact on comparisons of the demonstrated unreliability of portfolio scoring. A small number (14%) pointed to the wide variations in program implementation, such as the extent of focus on portfolio work and rules about task selection, outside help, and revision. A few (8 principals, or 11%) expressed concern about the adequacy of the small samples of students drawn by the state for scoring.

Twenty-three principals (30%) stated that comparisons among schools would be a misuse of portfolios. Twelve of these 23 (16%) stated that the appropriate function of portfolios is to provide in-depth measurement of

individual students rather than comparisons among schools. For example, one stated:

> **Publicizing the results defeats the purpose of portfolios, which is to be an individual assessment. They can't be standardized. If the state feels the need for such a measure, use a standardized test.**

We did not ask about the basis for principals' opinions of the proper functions of portfolios, but the wording of some of their answers suggests that some believe that the state's original goals for the program similarly stressed individual measurement rather than comparisons among groups. If so, that would, in our opinion, constitute a fundamental misunderstanding of the program's explicitly stated goals. Further interviews would be needed to determine the basis for these principals' views.

CHAPTER 4: THE RELIABILITY OF MATHEMATICS PORTFOLIO SCORES

In the 1991-92 school year, the scoring of Vermont's portfolios was unreliable in both subjects (mathematics and writing) and grades. The reliability of scoring was low enough to preclude most intended uses of the portfolio scores. Extensive analysis of the reliability of the 1991-92 portfolio scores is reported in Koretz, Stecher, Klein, McCaffrey, and Deibert (1993).

A central question in our studies was therefore whether the reliability of portfolio scoring improved in the 1992-93 school year. In mathematics, reliability did improve substantially.

This chapter discusses the reliability of mathematics scores in 1993 and compares it to the previous year, focusing primarily on rater reliability—that is, the consistency of ratings. The broader question of the reliability of scores is addressed by a generalizability analysis.[14]

We investigated rater consistency and score reliability at three levels. The first level was the scores assigned to each piece in the portfolio on each of the seven dimensions. The second was dimension-level scores. These were obtained by computing the student's mean score across all pieces within a dimension. The final level was a single overall mean score for each portfolio across all pieces and dimensions (i.e., the mean of the 7 dimension scores).

Interrater Correlations

We used Spearman rank-order correlations to examine rater consistency. We chose Spearman coefficients for this purpose because the 4-point scales on which pieces were graded were viewed by Vermont's teachers as uneven steps along a continuum rather than equal intervals, theoretically making the more common Pearson correlations inappropriate.[15] This correlational analysis found that in 1992, raters did not agree highly with each other in the score they assigned to a piece on a dimension or in the mean score they gave to the whole

---

[14] The material in this chapter has been submitted for journal publication. We are indebted to Lee Cronbach for detailed comments on the generalizability analyses presented here and in Chapter 5.

[15] In practice, using Pearson correlations yielded nearly identical results.

portfolio on that dimension (i.e., averaged across all the pieces in the portfolio). This was true for all 7 dimensions. In 1992, the correlations between raters at the piece level ranged from .30 to .41 at Grade 4 and from .31 to .47 at Grade 8. No dimension score (when averaged over pieces) consistently stood out from the rest as having an unusually high or low degree of rater agreement (see Appendix A). There was a similarly narrow range in 1993. Consequently, to simplify the discussion that follows, the piece- and dimension-level correlations reported in the tables below are the averages of the 7 dimension correlations.[16]

In 1992, piece-level correlations were below .40, and dimension-level correlations were only trivially better (Tables 4.1, 4.2). Total-score correlations were considerably higher but did not exceed .60 (Table 4.3). Rater reliability increased appreciably at all three levels (piece, dimension, and total scores) in 1993 (Tables 4.1–4.3) but only reached a .70 or better for total scores. The improvement between 1992 and 1993 (which was evident at the levels) may have stemmed from the greater control over rater training and grading in 1993.

Dependencies

A rater's evaluation of one piece on one dimension did not appear to be independent of that rater's assessment of another piece either on the same dimension or on a different dimension. There was less evidence of such dependency in 1993 than in 1992, but in both years, there was more agreement within than between raters.

Dependencies are suggested by higher correlations among pieces in a portfolio when these pieces are graded by the same rater than when they are graded by different raters. For example, in 1992 at Grade 4, the average correlation between two pieces in a portfolio on a given dimension was .27 when the same rater assigned both scores, but only .13 when one rater assigned the score to one piece and a different rater assigned the score to the other piece. In three of the four cohorts studied, the correlation between the scores assigned to different pieces on different dimensions when graded by the same rater was actually slightly higher than the correlation obtained when different raters graded these same pieces on the same dimension.

---

[16] Correlations were transformed to *z*-scores before averaging.

Table 4.1

Piece-Level Correlations Between Raters, Mathematics (Within-Dimension Correlations Averaged Across Dimensions)

	1991-92	1992-93
Grade 4	.34	.46
Grade 8	.37	.50

Table 4.2

Dimension-Level Correlations Between Raters, Mathematics (Within-Dimension Correlations Averaged Across Dimensions)

	1991-92	1992-93
Grade 4	.42	.57
Grade 8	.38	.65

***Note*. Dimension-level scores were created by averaging scores across pieces for each dimension. These composite scores were then correlated across raters, and the resulting correlations were averaged across dimensions.**

Table 4.3

Total Score Correlations Between Raters, Mathematics (Combining All Dimensions and Pieces)

	1991-92	1992-93
Grade 4	.60	.72
Grade 8	.53	.79

Dependencies are even more evident when the analysis is conducted at the dimension level. The correlation between scores on two dimensions was much higher when those scores were assigned by the same rater than when they were assigned by different raters (Table 4.4).

Table 4.4

Mean Correlation Between Two Dimensions When the Scores on These Dimensions Are Assigned by the Same Versus Different Raters

	1992		1993	
	Grade 4	Grade 8	Grade 4	Grade 8
Same rater	.35	.36	.50	.53
Different raters	.20	.20	.39	.44

Note. The score a rater assigned to a portfolio on a given dimension is the mean of the scores that rater assigned to all the pieces in this portfolio on that dimension.

To sum up, the results above suggest that the score a rater assigned to a piece on one dimension was not independent of the scores that rater assigned to other pieces in the student's portfolio on that or some other dimension. Similarly, the score a rater assigned to a piece on one dimension did not appear to be independent of the scores that rater assigned to that same piece on other dimensions.

There are many ways in which dependencies can arise. For example, there may be some characteristic of a student's work (such as "neatness") that cuts across all the pieces in a portfolio and which affects raters differently. When this characteristic is present, some raters may tend to give higher grades while others assign lower grades or are unaffected by its presence. Dependencies also may arise if a rater's assessment of one piece affects that rater's evaluation of the other pieces in the portfolio. This could occur because a rater graded all the pieces in a portfolio on all dimensions before grading another portfolio.

Intrarater Agreement

We also examined whether differences between raters in their evaluations of a student's portfolio were due mainly to random versus systematic factors. If differences stemmed from random factors, such as fatigue, then the degree of agreement within a rater on different days would be no higher than it was between raters. Conversely, if raters agreed more with themselves than with each other regardless of when they evaluated a portfolio,

then it would suggest systematic factors were at work, such as raters having consistent but idiosyncratic views about piece quality. If that was the case, then it would suggest that more effort was needed in rater calibration to ensure that all raters applied the same criteria and in the same way when they graded a piece.

This issue was investigated in conjunction with the 1993 rescoring activities by having a rater on the fifth day of the scoring session regrade two of the portfolios that same rater graded on the morning of the third day.[17] There were 71 Grade 4 and 53 Grade 8 portfolios in this intrarater study. Although the raters on the fifth day could not see the scores they assigned to a portfolio on the third day, they nevertheless were more consistent with themselves than they were with each other in their evaluations of the relative quality of a student's work (Table 4.5). These results suggest that further training of raters may improve the degree of agreement between them and thereby increase the overall reliability of portfolio scores.

Score Reliability

We used a generalizability analysis to assess how much of a piece's score on a dimension was a function of systematic differences in students, raters, pieces within a portfolio, and interactions of these factors. Although this analysis was done separately for each dimension, the percentage of variance

Table 4.5

Mean Spearman Rank Order Correlations Within and Between Raters in 1993

	Grade 4		Grade 8	
	Within raters	Between raters	Within raters	Between raters
Piece	.61	.46	.70	.50
Dimension	.67	.57	.81	.65
Total score	.79	.72	.92	.79

[17] A number of raters did not complete rescoring of a second portfolio. Because this was not known until long after the scoring session, we were not able to ascertain the reason.

attributable to a factor on one dimension was very similar to its percentage on other dimensions. Thus, to simplify the discussion that follows, we present just the averages of the percentages across dimensions (Appendix B contains the data for each dimension separately).

Reliability of Dimension-Level Scores

Table 4.6 shows that in 1992, consistent differences among students accounted for only about 15% of the variance in dimension scores. The rest of the variance was due to systematic or random error. About 15% of the variance was due to a combination of systematic differences between raters (i.e., one being more lenient than another) and the "Student X Rater" interaction (i.e., disagreement between raters in their assessment of the relative quality of a student's work). About one-fourth of the variance stemmed from students having a relatively high score on one piece but a moderate or low score on other pieces (i.e., there was a large "Student X Piece" interaction). Almost half the variance was due to residual error, which included interactions between raters and pieces. As a consequence of these effects, the reliability of a student's score on a typical dimension was only about .33 for a 7-piece portfolio read once.

The picture improved somewhat in 1993. Compared to the previous year, there was a reduction in variance due to raters and to the interaction between students and raters. In contrast, the Student X Piece interaction was larger in

Table 4.6

Percentage of Variance on a Typical Dimension That Was Attributable to Various Factors

	1992		1993	
Source of variance	Grade 4	Grade 8	Grade 4	Grade 8
Students	14	12	15	21
Raters	9	7	2	2
Students X Raters	6	8	5	4
Students X Pieces	23	27	31	32
Residual	48	46	47	41

1993 than in 1992. Only in Grade 8 was there a noticeable increase in the percentage of variance due to systematic differences among students, but the size of this component must still be considered small. The reliability for a 7-piece portfolio read once was still only .45 at Grade 4 and .55 at Grade 8.

The Effect of Additional Raters and Pieces on the Reliability of Dimension Scores

The g-study results were used to estimate the reliability of a student's score on a dimension as a function of two factors: the number of pieces in the portfolio and how many independent raters graded each portfolio. In this context, reliability is defined as the correlation between two scores on the same dimension where one score is based on one set of pieces graded by one rater (or set of raters) and the other score is based on a different set of pieces from the same student graded by a different rater (or set of raters). This interpretation assumes pieces are assigned randomly to the two portfolios from a hypothetical population of best pieces.

A reliability of .90 or higher is generally considered necessary for reporting results for individual students. Although the Vermont Department of Education does not intend to report portfolio scores for individual students, individual schools or districts may choose to, and other state programs may attempt to use portfolios in this manner. Thus it is useful to compare portfolios to that standard. None of the dimension scores came close to that level. This finding led us to examine the likely effects of three strategies for improving reliability—increasing the number of raters per portfolio, increasing the number of pieces per portfolio, and having each piece evaluated by a separate rater. None of these strategies are especially effective.

Table 4.7 shows that using more than two raters per portfolio increases the reliability of a student's dimension score only slightly. This occurred because systematic differences in mean scores between raters account for such a small portion of the variance.

Because the Student X Piece interaction is so large relative to the variance due to Students, even a substantial increase in the number of pieces in a portfolio will not produce a highly reliable dimension score. For example, doubling the number of pieces is estimated to increase the reliability of a 1993 fourth grader's score on a typical dimension from .45 to .54, and an eighth

Table 4.7

Estimated Reliability of a Student's Score on a Dimension in a 7-Piece Portfolio Graded by 1, 2, or 3 Raters

Number of raters per portfolio	1992		1993	
	Grade 4	Grade 8	Grade 4	Grade 8
1	.35	.32	.45	.55
2	.47	.45	.56	.65
3	.55	.52	.61	.70

grader's score from .55 to .64. These estimates assume the second 7 pieces behave the same way as the first 7, but this may not happen because the first 7 are supposedly the best examples of the student's work.

Using a different rater for every piece in a portfolio is analogous to having a separate rater for each question on an essay test. This strategy eliminates the dependencies between pieces discussed above, which in turn increases the reliability of a dimension score. However, the size of the increase would be small. For example, our g-study analysis indicates that if every Grade 4 student in 1993 had two portfolios (with 7 pieces in each one), then the correlation between the scores on the first and second portfolios on a typical dimension would be about .45 if one rater graded all the pieces in a portfolio versus .54 if each piece was graded by a different rater. The corresponding increase in Grade 8 would be from .55 to .64. Hence, using a separate rater for each piece within a portfolio would produce about the same small increase in the reliability of a dimension score as having each portfolio graded twice or doubling the number of pieces graded once by a single rater. Using a separate rater for each piece also complicates the logistics of the grading process.

The Effect of Additional Raters and Pieces on the Reliability of Total Scores

A student's total portfolio score is the average of all the scores assigned to that portfolio by a single rater (i.e., the mean across all dimensions and pieces). In the context of g-theory, the reliability of this score is the correlation between the total scores on different portfolios produced by the same student when these portfolios are graded by different raters.

According to our g-study for 1992, total score reliability would be .49 in Grade 4 and .45 in Grade 8. The values for 1993 were .63 and .71 (see Appendix B). These estimates are for an average length portfolio (i.e., 6 pieces graded by a single rater). A portfolio would have to contain a very large number of pieces before total scores would approach an acceptable level of reliability. Indeed, given the 1993 data, even the .85 level cannot be reached at either grade level unless each portfolio has over 25 pieces and each piece is graded by two or more raters (i.e., over three hours of grading time per portfolio).

Discussion

The degree of agreement among Vermont's portfolio raters was much lower than among raters in studies with other types of constructed response measures. Four factors appear to contribute to this situation: (a) dependencies arising out of a rater grading all the pieces in a portfolio on all dimensions before grading the next portfolio, (b) systematic differences among raters in how they interpret and apply the scoring rubrics, (c) the nature of these rubrics, and (d) the tremendous diversity (lack of standardization) of tasks across portfolios.

The dependency (or "halo" effect) problem can be ameliorated by having a different rater evaluate each piece (or each dimension), but this strategy greatly complicates the scoring process, and it is unlikely to appreciably increase the reliability of a student's score. Doubling the number of pieces in a portfolio or doubling the number of times each piece is graded produces about the same modest increase in score reliability as having each piece graded by a different rater.

The higher intrarater than interrater correlations suggest there are systematic differences among raters in their application of the scoring criteria. While it is an open question whether further training can substantially reduce or eliminate these idiosyncratic views, we noticed that interrater agreement in mathematics increased somewhat between 1992 and 1993, which corresponded with the change from multiple scoring sites to a single site and the implementation of more standardized rater training and calibration procedures. This extended training and calibration was not just a matter of initial, prerating refreshers (all the raters had been trained

previously). It meant that every rater heard the same discussions during the many calibration sessions.

Many performance assessment programs minimize disagreements among raters by employing scoring rubrics closely tailored to individual problems. This sort of scoring is not feasible in Vermont, because teachers and students are free to select their own problems, and raters do not know in advance what problems will appear in portfolios. In writing, high rater agreement rates have been attained with portfolios by applying rubrics specific to genres rather than to specific pieces (Gentile, 1992). Whether such an approach is feasible in subjects such as mathematics remains unclear, but at present, Vermont lacks a clear enough typology of mathematical problems to permit this approach.

CHAPTER 5: THE RELIABILITY OF WRITING PORTFOLIOS SCORES

In contrast to mathematics, the scoring of writing portfolios did not improve appreciably between 1992 and 1993. Because the 1992 results are very similar to those from 1992, and because the 1992 results have been described in detail elsewhere (Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993), we present only a brief summary of findings here.

As noted earlier, each portfolio generated two sets of scores, one for the best piece and the other for the remainder of the portfolio, called here the "rest" score. Both sets comprised scores on five 4-point scales.

Simple correlations between raters improved only slightly in 1993. At the level of individual dimensions and parts of the portfolio, correlations were .45 or less (Table 5.1). Dimension-level correlations were roughly .50 (see Table 5.2). Results for the "rest" scores were similar and are not presented).

Table 5.1

Piece-Level Correlations Between Raters, Best Pieces (Within-Dimension Correlations Averaged Across Dimensions)

	1991-92	1992-93
Grade 4	.35	.40
Grade 8	.42	.45

Table 5.2

Dimension-Level Correlations Between Raters (Within-Dimension Correlations Averaged Across Dimensions)

	1991-92	1992-93
Grade 4	.39	.46
Grade 8	.49	.52

***Note*. Dimension-level scores were created by averaging scores across the best-piece and "rest" score for each dimension. These composite scores were then correlated across raters, and the resulting correlations were averaged across dimensions.**

Reliability can be increased by creating total scores across both parts and all dimensions—a greater simplification of scores than the state intended—but even these total scores showed interrater reliabilities around .60 (Table 5.3).

We conducted a generalizability analysis of the 1993 writing scores to parallel that reported for 1992 (Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). This too showed very little change from 1992 to 1993 (Table 5.4). In 1993 as in 1992, differences among students accounted for only about one-third of the total variance of portfolio scores, while systematic and random error account for the rest.

Finally, we examined the consistency of scores *within* raters with the consistency *across* raters. Each rater was given a small number of portfolios to score a second time, two days after they had scored those portfolios the first time. The correlations between the first and second scores by each rater were compared to the correlations across raters, when different raters provided each of the two scores.

The intrarater correlations were higher than the interrater correlations—markedly so in Grade 4 (Table 5.5). There appear to be two plausible explanations of this pattern. It could arise from raters remembering the scores they assigned the first time. We were unable to poll raters systematically in this regard, but informal comments by some indicated that they recognized the portfolios but did not recall their first scores. (Raters scored for approximately 16 hours over two days between their first and second ratings of the portfolios involved in this portion of the study.) The pattern could also reflect systematic differences among raters that were not eliminated by training. That is, raters may differ in their interpretation of the scoring criteria.

Table 5.3

Total Score Correlations Between Raters, (Combining All Dimensions and Both Parts)

	1991-92	1992-93
Grade 4	.49	.56
Grade 8	.60	.63

Table 5.4

Variance Components as a Percent of Total Variance (Results Averaged Across Dimensions)

	1992		1993	
	4th	8th	4th	8th
Students	28	36	30	37
Raters	13	8	9	5
Students X Parts	7	7	10	8
Raters X Students	18	18	13	16
Raters X Parts	1	1	1	0
Residual	33	31	36	34

Table 5.5

Intrarater and Interrater Correlations (Results Averaged Across Dimensions)

	Within-rater	Interrater
Typical dimension, Best Piece		
Grade 4	.71	.40
Grade 8	.58	.45
Typical dimension, "Rest" score		
Grade 4	.60	.41
Grade 8	.52	.45
Total score, both parts and all dimensions		
Grade 4	.83	.56
Grade 8	.77	.63

APPENDIX A

Table A.1

Spearman Rank Order Correlations Between Raters at the Piece Level

	1992		1993	
Dimension	**Grade 4**	**Grade 8**	**Grade 4**	**Grade 8**
PS1-Understanding of Task	.30	.32	.46	.52
PS2-How: Procedures	.33	.36	.48	.50
PS3-Why: Decisions	.35	.37	.48	.49
PS4-What: Outcomes	.30	.31	.35	.36
C1-Language of Math	.34	.32	.43	.54
C2-Math Representations	.41	.47	.60	.60
C3-Presentation	.39	.45	.43	.52
Mean	.34	.37	.46	.50

Table A.2

Spearman Rank Order Correlations Between Raters at the Dimension Level

	1992		1993	
Dimension	**Grade 4**	**Grade 8**	**Grade 4**	**Grade 8**
PS1-Understanding of Task	.42	.38	.60	.69
PS2-How: Procedures	.48	.38	.61	.68
PS3-Why: Decisions	.48	.37	.63	.65
PS4-What: Outcomes	.43	.39	.44	.56
C1-Language of Math	.30	.28	.52	.70
C2-Math Representations	.36	.34	.63	.60
C3-Presentation	.51	.53	.58	.68
Mean	.42	.38	.57	.65

APPENDIX B

Two factors influence the score a student receives on a given dimension: the characteristics of the pieces in the portfolio and the rater who graded them. The interest is not in specific pieces or raters, but in the student's mathematics ability. A dimension score must therefore generalize from the observed score to the average score the person would receive on all pieces that might be included in a portfolio and raters who might grade them. To measure the generalizability of the dimension scores, we computed the proportion of the variance in dimension scores that was attributable to differences in student abilities, variation among the pieces in a given portfolio and the preferences and biases of the individual raters. The larger the proportion of the variance attributable to pieces and raters, the smaller the reliability of a dimension score and the less dependable or generalizable the measure.

Because each dimension is designed to measure a different ability (and arbitrarily chosen to represent a greater domain of mathematics skills), the dimensions do not represent a facet that must generalize. All analyses treated the dimensions as *fixed* and a separate analysis was conducted for each one. Furthermore, all results pertain only to these seven dimensions. The total portfolio score (i.e., the average of the seven dimension scores) is a measure of a student's overall mathematics ability. Hence, an analysis was also conducted using this mean score. In all analyses, the components of the variance among single readings of individual pieces were calculated. These variance components were used to estimate the reliability of portfolios with various numbers of pieces scored by specified numbers of raters.

In the sample of piece scores for a given dimension (or the average over dimensions), each score depends on three factors—the student and the two generalizable facets: rater and piece within the student. Thus, the score can be modeled as

$$y_{ijk} = \mu$$
$$+ \mu_i - \mu \quad \textbf{(student effect)}$$
$$+ \mu_j - \mu \quad \textbf{(rater effect)}$$
$$+ \mu_{ij} - \mu_i - \mu_j + \mu \quad \textbf{(student by rater effect)}$$
$$+ \mu_{ik} - \mu_i \quad \textbf{(piece within student effect)}$$
$$+ y_{ijk} - \mu_{ij} - \mu_{ik} + \mu_i + \mu \quad \textbf{(residual effect)}$$

where y_{ijk} is the score for the kth piece, (k = 1 to number of pieces included in the portfolio), from the ith student's portfolio (i = 1 to the number of students), graded by rater j (j = 1 to the number of raters). The procedure used for scoring student portfolios did not use a fully crossed design—many raters participated in the scoring, but only two of them scored a given portfolio. Thus, the data set is unbalanced, because all raters did not grade all portfolios.

The pieces in a portfolio are dependent on the student. No standard set of tasks was completed by all students, and pieces were randomly numbered within a portfolio. Hence, pieces are nested within the portfolio and represent a random sample of work from each student. For this reason, no piece component is included in the model.

Not only are pieces nested with student, but students are not fully crossed with raters. Each student's portfolio was scored by only two raters and raters scored portfolios from only a subset of students. Pairs of raters, however, were not nested within groups of students and this allows us to recover a rater effect. We treat our data as an unbalanced student by rater design (with pieces nested within student).

Because of the unbalanced nature of the data set, the traditional ANOVA-based estimates of components of variance were not available. The component estimates were found using the MIVQUE estimation procedure (Hartley, Rao, & LaMotte, 1978) because the large sample size made it impractical to use other ANOVA-based methods to estimate the variance components. MIVQUE produces unbiased, (locally) minimum variance estimates of the variance components.

Tables B.1 to B.4 contain the percentage of variance in the piece scores that is attributable to each effect on each individual dimension. Table B.5 contains the corresponding data for the total score on each piece.

The generalizability of the portfolio score for a single dimension is measured using the generalizability coefficient (Shavelson & Webb, 1991). The generalizability coefficient is approximately equal to the expected value (average) of the square of the correlation between the observed scores and the student's universe scores (Shavelson & Webb, 1991). It is also approximately equal to the correlation between observed scores on two analogous portfolio scores.

Under some general assumptions, the variance components yield estimates of the generalizability coefficient for any combination of parts and raters. The generalizability (reliability) coefficients, such as those given in Table 4.7, are calculated using the following formula

$$\frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2 / n_r + \hat{\sigma}_{sp}^2 / n_p + \hat{\sigma}_{srp.e}^2 / n_r n_p}$$

where $\hat{\sigma}_i^2$ i = s, sr, sp and srp.e denote the estimated variance components for students, student by rater, piece within student and residual effects respectively, and n_r and n_p denote the number raters scoring each portfolio and the number of pieces, respectively. The rater effect is included in the denominator of the formula, because the scoring procedure always includes several raters to score portfolios from the entire student population.[18]

The generalizability coefficients are unbiased as long as the variance components adequately capture the variability among scores from all pieces that might be included in a portfolio and all raters who might grade them. The observed variance among pieces within a portfolio might understate the

---

[18] Traditionally the rater effect would not be included in the denominator of the Generalizability Coefficient and our coefficient is often called the Index of Dependability, denoted φ (Shavelson & Webb, 1991). We include the rater effect in the denominator because the population of students will always be scored by several raters, and hence rater will always influence the relative standing of the student. This is consistent with the stated goal of the Generalizability Coefficient to account for all sources of error that influence the relative standing of individuals.

variance among all possible pieces because the selected pieces were chosen to be the student's best work. If a student chose additional pieces, they may be more variable. Furthermore, the selected pieces represent only a subset of all tasks a student could undertake. Pieces based on different tasks may differ from those currently included in the portfolios because a student's performance may vary with task. Finally, a single rater scored all the pieces in the portfolio. If contiguous scoring tends to understate the variability among pieces within the portfolio, then the observed variance underestimates the true variability that exists among all possible pieces. (The effects of contiguous scoring are not a problem as long as we consider only this scoring method. If, however, we use the observed variance components to calculate the reliability of scores where each piece in the portfolio is scored by a separate rater, then the estimated coefficient might be biased upward.)

Table B.1

Sources of Variance as Percent of Total Variance in a Piece-Level Score, by Dimension, Grade 4, 1992

	Dimension							
Source	**PS1**	**PS2**	**PS3**	**PS4**	**C1**	**C2**	**C3**	**Ave.**
Student	13	15	20	11	8	7	21	14
Rater	5	4	16	7	15	8	10	9
Student X Rater	7	5	3	9	6	6	4	6
Piece within Student	19	20	15	25	27	36	19	23
Residual Error	56	57	45	48	43	43	46	48
Total	100%	100%	100%	100%	100%	100%	100%	100%
Total Variance	0.39	0.58	0.77	0.19	0.49	0.63	0.71	0.54

Table B.2

Sources of Variance as Percent of Total Variance in a Piece-Level Score, by Dimension, Grade 8, 1992

	Dimension							
Source	**PS1**	**PS2**	**PS3**	**PS4**	**C1**	**C2**	**C3**	**Ave.**
Student	11	10	14	11	12	6	23	12
Rater	3	4	11	5	10	8	6	7
Student X Rater	10	9	10	1	11	4	9	8
Piece within Student	24	28	23	24	24	41	22	27
Residual Error	52	49	41	59	43	41	40	46
Total	100%	100%	100%	100%	100%	100%	100%	100%
Total Variance	0.37	0.54	0.79	0.14	0.50	0.69	0.81	0.55

Table B.3

Sources of Variance as Percent of Total Variance in a Piece-Level Score, by Dimension, Grade 4, 1993

	Dimension							
Source	**PS1**	**PS2**	**PS3**	**PS4**	**C1**	**C2**	**C3**	**Ave.**
Student	16	16	18	7	17	11	21	15
Rater	0	0	0	0	4	3	3	2
Student X Rater	4	4	7	6	5	2	7	5
Piece within Student	32	32	30	27	28	48	23	31
Residual Error	48	48	46	60	46	36	46	47
Total	100%	100%	100%	100%	100%	100%	100%	100%
Total Variance	0.36	0.47	0.50	0.10	0.35	0.60	0.63	0.43

Table B.4

Sources of Variance as Percent of Total Variance in a Piece-Level Score, by Dimension, Grade 8, 1993

	Dimension							
Source	**PS1**	**PS2**	**PS3**	**PS4**	**C1**	**C2**	**C3**	**Ave.**
Student	21	22	24	13	22	16	28	21
Rater	0	2	2	2	0	6	5	2
Student X Rater	5	3	5	2	4	6	3	4
Piece within Student	32	30	26	36	32	41	26	32
Residual Error	41	43	43	48	41	31	39	41
Total	100%	100%	100%	100%	100%	100%	100%	100%
Total Variance	0.39	0.69	0.80	0.28	0.52	0.64	0.68	0.54

Table B.5

Sources of Variance as Percent of Total Variance in Total Scores for a Piece

	1992		1993	
Source	**Grade 4**	**Grade 8**	**Grade 4**	**Grade 8**
Student	24	22	27	35
Rater	8	6	1	2
Student X Rater	6	11	5	5
Piece within Student	26	31	34	31
Residual Error	35	30	34	27
Total	100%	100%	100%	100%
Total Variance	0.25	0.26	0.20	0.26

REFERENCES

Gentile, C. (1992). ***Exploring new methods for collecting students' school-based writing: NAEP's 1990 portfolio study.*** **Washington, DC: National Center for Education Statistics.**

Hartley, H. O., Rao, J. N. K., & LaMotte, L. (1978). A simple synthesis-based method of variance component estimation. ***Biometrics, 34,*** **233-244.**

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). ***Interim report: The reliability of Vermont portfolio scores in the 1992-93 school year*** **(RAND, RP-260). Santa Monica, CA: RAND. (Reprinted from CSE Technical Report 370, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing, December.)**

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (in press). The Vermont Portfolio Assessment Program: Findings and implications. ***Educational Measurement: Issues and Practice, 13*****(3).**

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). ***Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience*** **(RAND, RP-259). Santa Monica, CA: RAND. (Reprinted from CSE Technical Report 371, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing, December.)**

Mills, R. P., & Brewer, W. R. (1988). ***Working together to show results: An approach to school accountability in Vermont.*** **Monpelier: Vermont Department of Education, October 18/November 10.**

Shavelson, R. J., & Webb, N. M. (1991). ***Generalizability theory: A primer.*** **Newbury Park, CA: Sage Publications.**

Stecher, B. M., & Hamilton, E. G. (1994, April). ***Portfolio assessment in Vermont, 1992-93: The teachers' perspective on implementation and impact.*** **Paper presented at the annual meeting of the American Educational Research Association, New Orleans.**